Unit21.ai

Checklist

The agentic AI vendor evaluation checklist for financial crime

How to use this checklist

Every vendor pitching a financial crime product now claims some version of "agentic AI." Very few agree on what that actually means, and fewer still can back the claim up under direct questioning. This checklist gives you one evaluation framework to bring into any vendor conversation, so you can tell a platform that has genuinely automated the work from one that has repackaged an existing model with new marketing.

This checklist has four parts

01 A universal set of questions and red flags that apply to any agentic AI workflow

02 A supplement for each of five specific use cases: AML investigation, fraud prevention, sanctions screening, entity onboarding, and network analysis/consortium

03 A practical next steps for running your evaluation

04 A self-scoring scorecard to track vendors against

Bring the relevant sections into your next vendor call, RFP, or POC scoping conversation. A vendor can score a Level 3 on one use case and a Level 1 on another; score each workflow separately.

Section 1

Universal questions (any use case)

Design

- [ ] Can you configure which tasks the agent performs per queue, without re-engineering the whole agent?
- [ ] Can your own team add or remove tasks without an engineering dependency on the vendor?

Context & Information

- [ ] Does the agent generate deterministic code (SQL/Python) to query data, rather than having the LLM interpret raw data directly?
- [ ] Can the vendor walk you through how they scope context for a specific task — what's included, what's excluded, and why?

Control & Oversight

- [ ] Can autonomy be configured per queue and per risk tier, not just globally?
- [ ] What happens when the agent hits a low-confidence result — does it escalate, or guess?

Tool & Model Fit

- [ ] How many LLMs does the platform use, and how is the right one selected per task?
- [ ] What happens to your workflow when a model's benchmark performance changes?

Feedback & Evaluation

- [ ] Does the vendor have evaluation datasets for each task type, sourced from real human-reviewed decisions rather than synthetic data?
- [ ] Can the AI agent QA a human analyst's work, and can a human QA the AI's, in both directions?

Governance & Safety

- [ ] How does the vendor's AI system align with EU AI Act requirements today?
- [ ] Can they produce an audit trail with a full reasoning chain and evidence citations, not just a final decision?

Red Flags

- ✔ The vendor can't explain how context is scoped per task.
- ✔ No evaluation datasets, or synthetic ones only.
- ✔ The "AI agent" is a chatbot over a data lake with no structured workflow execution.
- ✔ No configurable autonomy ladder — it's all-or-nothing automation.
- ✔ No accuracy metrics benchmarked against human analysts.
- ✔ Audit trails show decisions but not the reasoning or evidence behind them.
- ✔ The system can't generate deterministic code for data queries.
- ✔ No feedback loop from human corrections back into the agent.

Section 2

Use-Case Supplements

AML Investigation

- [ ] How many specialized sub-agents run per alert — and do they run in parallel or in sequence?
- [ ] Does the agent generate deterministic code to query transaction data, or does the LLM read it directly?
- [ ] What's the source of the evaluation dataset — real analyst decisions, or synthetic data?
- [ ] Can autonomy be configured per queue, process, team, or jurisdiction?
- [ ] Does the agent draft SAR-ready narratives, or does it claim to file SARs outright? These are different claims — confirm which.

Fraud Prevention

- [ ] Can your fraud team build custom investigation tasks without an engineering ticket?
- [ ] Does every agent run produce three outputs: a recommended action, a regulator-ready narrative, and a transparent work log?
- [ ] Can autonomy be configured differently per fraud type (e.g., ACH vs. card vs. account takeover)?
- [ ] Does the agent adapt to and help identify new fraud typologies, or only detect known patterns?

Sanctions Screening

- [ ] Can the agent distinguish sanctions screening (binary list check) from PEP screening (relationship analysis)?
- [ ] What's the auto-close rate for clear non-matches, and what accuracy backs that rate?
- [ ] How quickly does the system incorporate new sanctions list designations?
- [ ] Is this a specialized sanctions agent, or a general-purpose agent running sanctions alongside unrelated tasks?

Entity Onboarding (KYC/KYB)

- [ ] Can the agent consume and correctly interpret onboarding data from multiple source systems?
- [ ] Does it integrate with customer risk rating, including CDD vs. EDD determination?
- [ ] How does it resolve conflicting data from two or more sources?
- [ ] Does it collect incremental data itself, or stop and hand back to a human at the first gap?

Network Analysis & Consortium Intelligence

- [ ] Does the consortium data include investigation outcomes and typology classifications, or just binary good/bad flags?
- [ ] Does the network agent see cross-institutional signals, or only your institution's own data?
- [ ] What's the diversity of the consortium — cross-vertical and cross-typology, or single-sector?
- [ ] Can the agent proactively flag entities from cross-network intelligence before they transact on your platform?

Section 3

Steps to Take

Step 1

Classify every vendor claim against the four-stage maturity spectrum before taking it at face value:

- ☐ Level 1, Rule-based: fixed logic; no learning, adaptation, or autonomous execution.
- ☐ Level 2, AI-assisted: the AI surfaces information or suggestions; a human does the actual work.
- ☐ Level 3, Agentic with oversight: the AI executes tasks autonomously within guardrails, and a human reviews before or after.
- ☐ Level 4, Autonomous agentic: the AI acts and adapts without per-task human review, within defined risk boundaries.

Step 2

- ☐ Demand transparency on evaluation datasets — ask what they are, where they come from, and what accuracy threshold must be met before deployment.

Step 3

- ☐ Test context engineering claims directly — ask the vendor to demonstrate, live, how they scope context for one specific task.

Step 4

- ☐ Start with one workflow, not a platform overhaul — validate accuracy in parallel with human review before expanding to others.

Step 5

- ☐ Configure autonomy progressively — start conservative, build trust, then increase autonomy as accuracy holds.

Step 6

- ☐ Build your accountability framework now — random sampling, dual-direction QA, audit trail review — don't wait until after go-live.

Step 7

- ☐ Move with appropriate urgency. Most of what's in this checklist is a solved problem today for the vendors that are actually at Level 3+; governance should enable speed, not block it.

Section 4

Self-Scoring Scorecard

Score your current or prospective vendor on a scale of 1–4 for each dimension, specifically for the workflow you're evaluating. Repeat this table separately for each use case — a vendor can score Level 3 in AML and Level 1 in sanctions.

Dimension	What "good" looks like at Level 3+	Your vendor's score (1-4)
Design	Tasks are configurable per queue; your team can add tasks without an engineering dependency	
Information & context	Context is engineered per task, not dumped in wholesale; queries are deterministic, not LLM-interpreted	
Control & oversight	Autonomy is configurable per queue/risk tier; low-confidence results escalate rather than guess	
Tool & resource fit	Multiple models orchestrated per task; latency-tolerant design	
Feedback & integration	Real evaluation datasets from human-reviewed decisions; dual-direction QA (AI checks humans, humans check AI)	
Governance & safety	Least-privilege execution; deterministic fallback when uncertain; current EU AI Act compliance in production	

The vendors worth signing with will welcome this level of scrutiny. For the rest, this is exactly the diligence that catches a mismatch before it costs you a budget cycle and a year of implementation, not after.

Next Steps

Read the framework this checklist came from

This checklist is one extract from a larger piece of research: Chartis Research's report on the agentic maturity spectrum in financial crime — an independent framework for what agentic AI means across AML, fraud, sanctions, onboarding, and network analysis.

Inside the Report

4

Maturity levels defined by Chartis Research

5

FinCrime use cases analyzed in depth

7

Evaluation dimensions buyers should test

Beyond this Checklist

- ☐ Chartis' precise definition of agentic AI — and the four things it explicitly is not
- ☐ A discovery-call script contrasting the Level 2 and Level 3+ answer to each question, side by side
- ☐ A level-by-level breakdown of all five use cases, so you know which workflows a platform actually serves at Level 3+
- ☐ The four pillars of AI reliability, and how to assess each one in production rather than on a roadmap
- ☐ Chartis' independent evaluation of Unit21 against every dimension of the framework

Cut through the noise. Read the full Chartis report.

unit21.ai/agentic-ai-maturity